International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(6s) 2026 pp. 1019-1029
https://doi.org/10.70917/ijcisim-2026-3001
Article

Deep-Learning-Enhanced Entity Resolution for Cyber Threat Intelligence Using OSINT Sources

Ramesh Kumar Sharma[1], Dharmendra Kumar Singh[2], Rajesh P. Barnwal[3]
[1]Department of CSE, BIT Sindri, Dhanbad–828123, Affiliated to (JUT), Ranchi–834010, Jharkhand, India
Email: rameshs.rs.cse.21@bitsindri.ac.in , sharmarameshdhn@gmail.com
[2]Jharkhand University of Technology (JUT), Ranchi, Jharkhand, India
Email: dksingh.bits@gmail.com
[3]Information Technology Group, CSIR–Central Mechanical Engineering Research Institute (CSIR-CMERI), Durgapur, India; Associate Professor (Hon.), AcSIR, India
Email: rbarnwal.cmeri@csir.res.in

Abstract: Cyber Threat Intelligence (CTI) relies on open-source intelligence (OSINT) more as a means to detect, link and assign malicious cyber activity. Another ongoing difficulty with OSINT-based CTI is entity resolution, in which heterogeneous mentions of the same real-world entity, e.g. IP addresses, malware families, and threat actor alias, need to be correctly and reliably identified across heterogeneous and noisy data sources. This is specifically challenging as naming conventions are inconsistent and when one writes in more than one language, and when rivals intentionally misname it. Conventional, rule based, and shallow learning methods are not very strong in these conditions. This paper suggests an entity resolution framework of CTI based on deep-learning to produce context-aware representations of cyber threat entities of unstructured OSINT data using transformer-based architectures. This framework defines entity resolution as a similarity learning and clustering problem with the use of multilingual embeddings and graph-based entity linking. Curated OSINT data experimental testing shows that the proposed solution performs very well when comparing the obtained results with baseline techniques in terms of precision, recall, F1-score, and quality of clustering. The findings also demonstrate better resiliency to alias ambiguity and textual unstructured variations. On the whole, the paper has shown that entity resolution using transformers is a cornerstone to the right correlation of threats, campaign analysis, and automated pipelines of CTI.

Keywords: Cyber Threat Intelligence; Open-Source Intelligence (OSINT); Entity Resolution; Entity Disambiguation

1. Introduction

Cyber Threat Intelligence (CTI) has become a platform of the contemporary cybersecurity practice as it helps defend in a proactive way, provides response to the incident, and makes the critical strategic decisions. The main subject of CTI is to gather and examine the information related to the threat actors, attack methods, infrastructure, and malware to predict and prevent cyber threats before they evolve into the big picture events [1], [2]. With the spread of cyber operations and their deep understated nature, intelligence-based security has turned to a reactive method of detecting intrusion to a predictive and contextual analysis of enemy activity.

The Open-Source Intelligence (OSINT) is an extremely important part of the CTI ecosystem because it is highly ubiquitous, real-time, and covers both technical and human-oriented threat indications. OSINT materials comprise social networks, shadowy forums, dark markets, paste archives, hack disclosure, malware archives, and white hat research findings [3]. Such sources are frequently the ones that give early warning of an impending threat, like compromised credentials, evidence of malware development (or reuse patterns of the infrastructure). Nonetheless, the uncontrolled and unfiltered quality of OSINT presents significant sophistication in its analysis as well

Entity resolution, which is the process of determining and uniting several representations of the same real-world object in heterogeneous data sets, is one of the most enduring problems in CTI based on OSINT. When applied to CTI, the entities consist of IP addresses, domain names, malware families, threat actor alias, social media accounts, and cryptographic hashes. Higher-level intelligence operations like threat actor attribution, campaign rebuilding, and building a cyber threat knowledge graph require correct entity resolution [4]. In the absence of good entity linking, intelligence is always fragmented and as a result, restrictions are made to the analysts in attempting to perceive coordinated activity or the long term plans of an adversary.

Manual methods of cybersecurity entity resolution have been mostly based on rule-of-thumb methods, pattern match, and string similarity methods. These techniques are adequate in the case of structured indicators (e.g., IP addresses or hashes), but they do not scale in situations where unstructured OSINT text and human-created identifiers are used [5]. The situation is aggravated by a number of domain-specific factors. To begin with, attackers have often assigned themselves various usernames on different platforms with different identities over time. Second, malware and crimes are named differently by vendors and communities, which means they do not have an equal label on the same piece of art [6]. Third, OSINT content can also be highly multilingual and code-mixed, a problem that increases the ineffectiveness of surface-based similarity scores further.

One of the major peculiarities of CTI OSINT is its adversarial character. As opposed to most of the traditional data integration areas, CTI data is always maliciously altered by individuals who have ill intentions of evading notice or miscatenying the analysts. Cyberscriminals could mimic other organizations, use established handles to generate counter-markers of a false identity, or introduce misleading details to online discussion rooms [7]. Consequently, entity resolution systems have to be run in a place where ambiguity and deception are not error conditions but natural occurrences. This poses a major constraint to the use of traditional entity matching methods.

New developments in the field of deep learning are providing new opportunities to overcome these challenges, especially with transformer-based architectures. Transformer models have shown outstanding results in natural language processing through the contextualized representations they learn to identify semantic meaning outside of the similarities among the tokens [8]. BERT and its variants processed by models can multimodally encode contextual signals, linguistic syntax and cross-language meanings, which is why it is suited to analyze messy and heterogeneous text using OSINT. The use of deep learning methods has already demonstrated the future potential of these tools in cybersecurity activities such as malware recognition, phishing, and intrusion recognition [9], [10].

Regardless of this development, there is little research to date on the use of transformer-based models to entity resolution in CTI OSINT. The vast majority of entity resolving research in the area of deep learning has been performed in the field of e-commerce, digital libraries, and general-purpose knowledge graphs where the data is relatively clean and non-adversarial [11]. All these cannot be assumed to be true in CTI environments where things change very fast and data quality is quite fluctuated. The result is an evident gap in the research in terms of creating entity resolution structures which utilize deep semantic representations without necessarily taking into consideration the working conditions of cyber threat intelligence.

The paper fills this gap with a recommendation of an entity resolution framework that enhances the CTI OSINT framework leveraging the power of transformer-based architectures to create contextualized representation of cyber threat entities. The proposed solution adds semantic context and multilingual understanding and similarity learning in order to enhance entity disambiguation of the various OSINT platforms. The value of this piece is its contribution in the development and progression of AI powered CTI processes, and how transformer-based entity resolution, as well as methods, can enhance automated intelligence pipelines and validate cyber threat analysis in a more accurate manner.

2. Background and Related Work

The increased use of Open-Source Intelligence (OSINT) has strongly influenced the Cyber Threat Intelligence (CTI) processes by allowing the mass surveillance of the activities of the adversaries that are no longer contained within the limited scope of the telemetry of the conventional perimeter. CTI systems based on OSINT processes are used to derive threat-relevant data on the heterogeneous sources, including social media, underground forums, databases of vulnerabilities, malware depositories, and security blogs [12]. These sources are good sources of early-warnings, but they are also noisy, redundant, ambiguous, and difficult to analyze automatically.

One of the central analytical operations in CTI is entity resolution which seeks to establish whether or not different observations are based on the same underlying cyber entity. Initially the methods of CTI entity resolution were limited to indicator centered processes that attempted to coerce structured entities e.g. IP addresses, domains and

indeed cryptographic hash to precisely match [13]. These methods work well in low-level association, but cannot work when the entities being correlated have an indirect representation in natural language, or aliases, or changing identifiers. To counter this weakness, graph-based and ontology-driven models were presented to model the entity-relationships to allow deeper contextual reasoning [14]. These models are however limited in scalability and tailoring since they rely mostly on manual curation and fixed rules.

As machine learning developed, scientists started to investigate the use of statistical and supervised learning as a way of entity disambiguation. Classifier based on feature, using the flexibility of string similarity, time proximity and similarity of co-occurrences, demonstrated intermediate gains compared to rule-based systems [15]. However, these approaches are faced by challenges in multilingual OSINT data as well as being highly susceptible to feature engineering decisions. In addition, they are also prone to adversarial manipulation, which is a characteristic characteristic of cyber threat settings.

The latest advances in deep learning and dealings with transformer architectures have transformed the prospects of entity resolution studies. Transformers have shown great capabilities of extracting semantic and contextual information of unstructured text, including cross-lingual representations [16]. Transformer models have been used with cybersecurity addressing tasks like malware description and threat report classification, as well as automated extraction of intelligence [17], [18]. In spite of these developments, transformers specifically used in entity resolution in CTI OSINT are underutilized. The literature largely applies generic-purpose entity matching frameworks, making no specific reference to the challenges unique to CTI (e.g., alias reuse, deceit, and quickly changing threat identities) [19].

Table 1 describes several of the important methods of entity resolution and their limitations in the CTI OSINT scenario. In general, the reviewed literature exposes the existing knowledge gap regarding domain-sensitive, deep-learning-based entity resolution systems with application to adversarial intelligence settings, which objectives the approach to the current study.

Table 1. *Summary of Entity Resolution Approaches in CTI and Related Domains*

Approach Type	Key Techniques	Strengths	Limitations in CTI OSINT
Rule-based / Heuristic	Exact matching, string similarity	Simple, interpretable	Fails with aliases, multilingual text
Graph / Ontology-based	Knowledge graphs, relationship modeling	Contextual reasoning	Manual curation, limited scalability
Classical ML	Feature-based classifiers	Improved flexibility	Sensitive to features, weak to deception
Deep Learning (Transformers)	Contextual embeddings, semantic similarity	Handles unstructured text	Limited CTI-specific adaptation

3. Research Gap and Problem Formulation

Although there has been an increased interest in the understanding of how Cyber Threat Intelligence (CTI) could be applied with the use of the artificial intelligence, there is still a lack of awareness regarding the use of entity resolution in the CTI mechanisms in an OSINT-based setting, especially in adversarial and multilingual environment. The current CTI systems mostly use a rule-based correlation (or shallow machine learning) approach where the identifiers are assumed to be constant and the naming schemes are consistent [20]. The latter are not true in the field of the actual OSINT where an attacker may have several alias identities, may recycle infrastructure when switching campaigns, and they may intentionally include deceptive signals to avoid attribution [21]. The existing methods as a result tend to generate incomplete or inaccurate entity connections constraining the accuracy of the resultant intelligence products.

Although recent works have investigated deep learning and text analysis utilizing transformers into cybersecurity, most are concentrating on classification or information extraction problems and not on cross-source entity disambiguation [22]. The general-purpose entity resolution systems designed to operate in non-adversarial settings are unable to consider the CTI-specific issues (high rate of entity change, low density of ground truth labels,

high cost of false attribution) [23]. Furthermore, code-mixed and multilingual OSINT data also contributes to deteriorating the performance of models which were trained on monolingual data or purified data.

The above limitations outline an apparent gap in research: there was a lack of a domain-conscious, transformer-based entity resolution framework specifically tailored towards CTI OSINT. In order to eliminate this gap, the entity resolution problem is defined as a contextual similarity learning problem with the objective to identify two or more heterogeneous OSINT mentions to be associated with the same underlying cyber entity. Formally, having a body of entity mentions obtained by multiple OSINT sources, one can be asked to learn an entity-entity mapping such that intra-entity similarity, and inter-entity ambiguity are maximized and minimized respectively positioning the mapping under adversarial conditions. Using this formulation, sound entity linking is made possible which makes it possible to precisely correlate and attribute threats, which is useful in operational CTI systems.

4. Proposed Methodology

This paper outlines a transformer-based entity resolution system to be used with OSINT-based Cyber Threat Intelligence (CTI), aimed at dealing with alias ambiguity, multilingual information and adversarial deception. It is a pipeline-based methodology which is based on data acquiring, entity representation, similarity learning, and entity linking implemented in a structured workflow.

4.1 OSINT Data Collection and Preprocessing

OSINT information is gathered through sources of heterogeneity such as social media posts, forums, security blogs, malware reports, and even disclosure of vulnerability. These sources include not only organized identifiers (e.g. IP addresses, domains) but also unstructured pieces of text (e.g. threat actor handles, malware names). Preprocessing entails noises, normalization of indicators, language detection and sentence-level segmentation. In order to deal with the multilingual and code-mixed data, the text is stored as it is and implemented by the multilingual transformer models [24].

4.2 Contextual Representation and Entity Extraction

Relevant named entities to CTI are derived, including, but not limited to, IP addresses, domains, malware families, and actor aliases, through a pattern-based sensitivity and a neural named entity recognition model [25]. All extracted entity mentions are then embedded with the help of a transformer-based encoder that captures the lexical form of an entity and textual context. Contextual embeddings allow the model to differentiate between entities with surface forms that are similar, but the operational meaning differs, which is very common in OSINT data [26].

4.3 Entity Matching and Similarity Learning

Entity resolution is presented as a similarity learning problem on the pair-wise basis. With two entity mentions, a Siamese or cross-encoder transformer architecture evaluates a similarity score which is a probability that the two mentions represent one real-world entity. Fine-tuning of the model is done with positive and negative pairs, which are labeled through curated datasets of CTI. The idea of contrastive loss is applied in order to maximize the similarity with intra-entity and minimizing similarity with inter-entities [27].

4.4 Clustering and Entity Linking

The similarity scores are bundled together to form an entity similarity graph with nodes being entity mentions and resulting edges depicting the learned similarity weights. Hierarchical or density-based methods of clustering based on graphs are followed to cluster the mentions into solved entities [28]. To minimize the occurrence of false linkages, or non-coincidental similarity, sources and temporal constraints are added to the similarity.

Table 2 provides an overview of the main steps of the suggested methodology and purposes. Altogether, this framework will combine strong contextual insights with CTI-specific limitations, which makes it possible to resolve entities more resiliently and on a larger scale in adversarial OSINT settings.

Table 2. *Proposed Transformer-Based Entity Resolution Framework*

Stage	**Technique Used**	**Purpose**
Data Collection	Multi-source OSINT ingestion	Capture diverse threat signals

Preprocessing	Normalization, language detection	Reduce noise, preserve context
Entity Representation	Transformer embeddings	Context-aware encoding
Similarity Learning	Siamese / cross-encoder models	Learn entity equivalence
Entity Linking	Graph-based clustering	Resolve cross-source entities

5. Experimental Setup

Here, this section outlines the datasets, baselines, metric evaluation and implementation information utilized by evaluating efficacy of the suggested transformer-based entity resolution model paired with OSINT-derived Cyber Threat Intelligence (CTI).

5.1 Dataset Description

Experiments are designed to test the performance of cross-platform entity resolution on a curated OSINT CTI dataset that has been developed based on several publicly available sources of intelligence, such as security blogs, malware analysis reports, back floor discussions, and social media posts about cyber threats. The data with heterogeneous mentions of entities include IP addresses, domains, malware family names, and alias names of threat actors. Ground truth labels are acquired with the help of manual annotation with the assistance of cross-referencing to the known repositories of CTI and reports provided by the vendors in accordance with the general intelligence validation best practices [29]. The last dataset will be split into training (70%), validation (15%), and testing (15%) where entity mentions in the same real-world entity are not cross-slicing splits.

5.2 Baseline Methods

The approach here is contrasted with strategies that are employed by representatives of the baselines in CTI studies and the entity resolution studies. These are (i) string matching with rules based on exact and fuzzy similarity, (ii) feature based machine learning classifiers based on lexical and contextual features, and (iii) non-contextual embedding such as TF-IDF and Word2Vec with cosine similarity [30], [31]. The baselines are a very rich point of reference to assess contributions of deep contextual representations.

5.3 Evaluation Metrics

The evaluation is made by typical entity resolution measurements, such as precision and recall and F1-score, which are calculated at the resolved-entity level. Moreover, the clustering quality is evaluated with help of pairwise F1 and adjusted Rand index (ARI) as the measures of the accuracy of entity grouping among the sources of OSINT [32]. These measures are especially appropriate to the CTI cases, where missed linkages and false associations can both be very important in terms of operational impacts.

5.4 Implementation Details

During the setup of the transformer-based encoder, a multilingual model that runs on the CTI data is pretrained and then narrowed to the CTI data. Contrastive loss is used to do training using mini-batch sampling of positive and negative entity pairs. Hyperparameters like learning rate, batch size, similarity thresholds, etc. will be chosen using validation performance. Any given experiment is done in a controlled scenario to ascertain its reproducibility.

Table 3 summarizes datasets and baseline methods, which were compared.

Figure 1 represents the experimental workflow, which marks the path of the OSINT data ingestion to evaluation.

Table 3. *Experimental Setup Overview*

Component	**Description**
OSINT Sources	Security reports, forums, social media, malware analyses
Entity Types	IPs, domains, malware names, actor aliases
Baselines	Rule-based, classical ML, non-contextual embeddings

Metrics	Precision, Recall, F1, Pairwise F1, ARI
Model	Multilingual transformer with contrastive learning

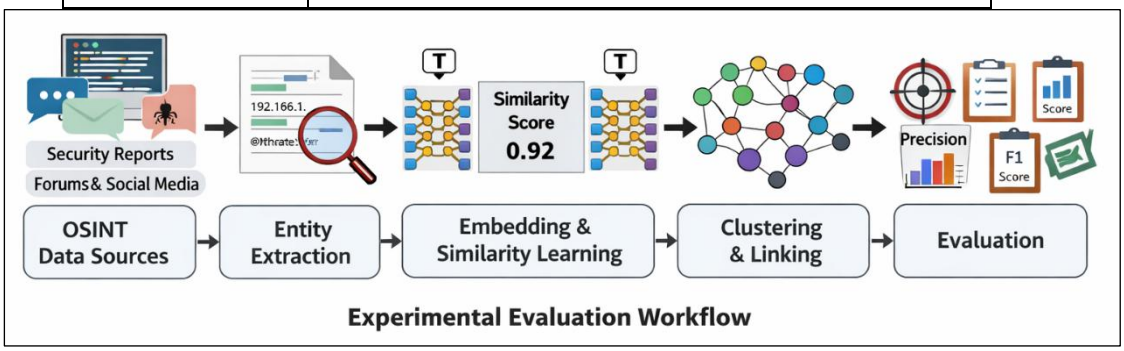

Figure 1. *Experimental Evaluation Workflow*

6. Results and Analysis

Further in this section, experimental findings achieved through the introduction of the transformer-based entity resolution framework are described and compared to those of the traditional methods of entity resolution. The testing is based on how well the resolutions of entities are correct, how well the quality of clustering is in the context of heterogeneous sources of OSINT, and how well the testing is robust.

6.1 Quantitative Performance Evaluation

The overall entity resolution effectiveness on various approaches is reported in Table 4 on precision, recall, and F1-score. The transformer-based method is significantly better than any of the baselines, with the highest F1-score. Specifically, it outperforms significantly rule-based and classical machine learning algorithms that have low recall owing to the impossibility of solving aliases and contextual variations. The non-contextual embedding techniques including TF-IDF and Word2Vec show moderate improvement with small caps highlighting the weakness to deal with the semantic ambiguity and multilingual OSINT content.

The findings demonstrate that contextual embeddings of transformer models promote models to capture semantic equivalency of entity mentioning to a great degree. This performance is particularly high in the case of threat actor alias names or malware names in which similarity at surface level is frequently deceptive.

Table 4. *Entity Resolution Performance Comparison*

Method	**Precision**	**Recall**	**F1-score**
Rule-based Matching	0.71	0.58	0.64
Classical ML	0.75	0.63	0.68
TF–IDF + Cosine	0.78	0.69	0.73
Word2Vec + Cosine	0.80	0.72	0.76
Proposed Transformer Model	**0.88**	**0.83**	**0.85**

6.2 Quality Analysis in Cluster

In addition to the accuracy of resolution at the pairwise level, the effectiveness of clustering is essential in CTI as resolved entities are usually exploited to build threat intelligence knowledge graphs. Table 5 includes the results of the evaluation of clustering, in terms of pairwise F1 and Adjusted Rand Index (ARI). The offered approach proves to be better in clustering coherence, meaning there are less erroneous mergers and splits of clusters of entities. This can be noted especially in adversarial CTI systems, in which attribution error can occur due to false linkage.

Table 5. *Clustering Performance Evaluation*

Method	Pairwise F1	ARI
TF–IDF + Clustering	0.68	0.54
Word2Vec + Clustering	0.72	0.60
Proposed Transformer Model	**0.81**	**0.71**

6.3 Cross-Source and Entity-type Analysis

The performance of the various methods in comparison to their respective entity types (see IP addresses, malware names and threat actor aliases) is shown in Figure 2. Although all the methods do reasonably well in structured indicators like IP addresses, major performance differences are realized between the unstructured entities. The suggested plan has the most significant difference in gains of the threat actor aliases, which emphasizes the significance of contextual knowledge in the process of resolving the alias.

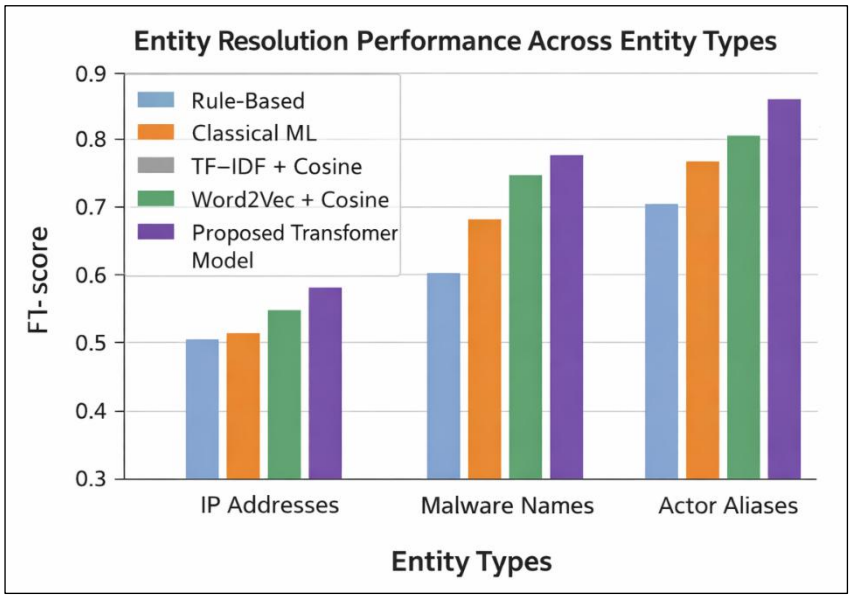

Figure 2. *Entity Resolution Performance Across Entity Types*

Figure 3 is a qualitative view of the resolved cluster of entities produced by the proposed model. As shown in the figure, the intra-cluster similarity is higher and more pronounced and the separation among distinct entities is more pronounced than with the baseline methods, which also validates the quantitative results of clustering.

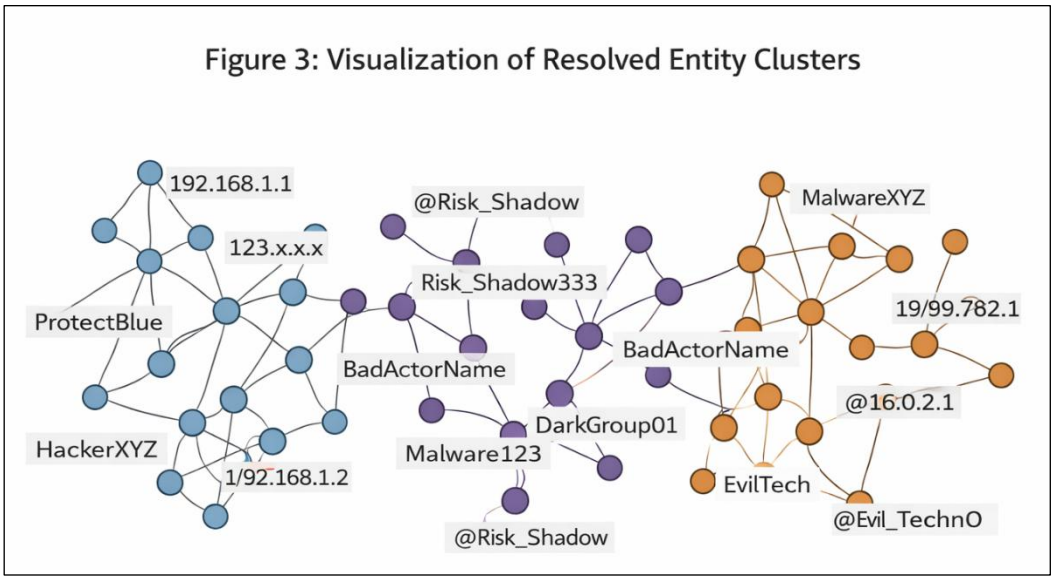

Figure 3. *Visualization of Resolved Entity Clusters*

6.4 Discussion of the Results

In sum, it is established that contextual representations based on transformers significantly enhance entity resolving in CTI based on OSINT. The suggested framework would not only increase the accuracy but offer more resilience to the ambiguity of Aliases, inconsistent naming, and multilingual text. This enhances the operational CTI systems directly to enable more accurate threat correlation, campaign tracking and intelligence-driven decision-making processes.

7. Discussion

The findings of the experiment prove that entity resolution using a transformer offers a significant enhancement compared to conventional methods used in OSINT-based Cyber Threat Intelligence (CTI). The reported improvement of both the pairwise and clustering results obtained bode well on the power of context-driven semantic modeling of heterogeneous, noisy and adversarial intelligence data. The proposed framework is unlike rule-based and classical machine learning approaches in that the relationships between entity mentions in the latent form are quite well explained by surface-level similarity alone.

One of the main lessons of the findings is that the difference in the performance of the entities between different types is uneven, and it represents the reality of implementing CTI. Organized objects like IP addresses perform fairly well under all the means, because they are standardized, and are less unclear. Malware names, and threat actor names, on the other hand, are highly variable because of inconsistent naming systems, vendor-specific names, and the deliberate use of re-use of alias. These categories are the ones where the proposed transformer-based model attains maximum performance benefits, highlighting its capabilities to use contextual cues, including the co-occurrence entity, temporal and narrative descriptions, to resolve ambiguity [33].

The clustering analysis also indicates the operational usefulness of deep contextual embeddings. The high clustering coherence means that the given approach minimizes the false merges (between unrelated things) and false splits (when one and the same thing can be found in different clusters). It is especially important in the case of CTI where a misidentification of links among the entities may cause incorrect threat attribution, an incorrect campaign identification, or the underestimation of the adversary capabilities [34]. The fact that transformer-based similarity learning yields smaller and semantically valuable clusters is validated by visualizing their solved clusters, which allows constructing trustworthy CTI knowledge graphs.

Intelligence operations have significant implications of these findings. Better entity resolution will positively complement cross-platform threat correlation, allowing the analyst to more reliably monitor the activity of the adversaries in forums, social media, and technical reports. The ability facilitates more advanced intelligence activities like analysis of a campaign, profiling the actor and the ability to anticipate and detect coordinated malicious activity at an early stage [35]. In addition, the decrease in the human workload of manual disambiguation can also reduce

analyst workload to a large degree, and the human expertise can be used to conduct strategic assessment than to clean up data.

Nevertheless, there are also some practical considerations that can be outlined in the results. Models based on transformers implement a more expensive computational model than lightweight heuristics and this might be restrictive to execution in real-time, given the resource constrained environment. Also, similarity learning on functions that lack ground truth or that may be sensitive to it using labeled data is problematic. Such restrictions are indicative that hybrid methods, i.e. deep learning combined with analyst-in-the-loop validation and incremental learning, might provide a viable way out [36].

On the whole, this discussion substantiates that the idea that deep-learning-enhanced entity resolution is not simply a performance upgrade but a bottom-line player of scalable, error-free, and robust CTI systems. This level of tackling entity ambiguity provides a better and improved strength of the entire intelligence pipeline and improvement in the state of automated cyber threat analysis.

8. Cyber Threat Intelligence Applications and Implications

The suggested entity resolution with deep-learning, has a strong practical use and implication on operational Cyber Threat Intelligence (CTI). The team framework enhances the precision and strength of industry linking between heterogeneous OSINT sources by enhancing various fundamental CTI operations, which rely on trustful identity discovery and situational correspondence.

Threat actor profiling and attribution is one of the major uses. Correctly enumerating aliases, handles and malware variant names allows analysts to build a more coherent profile in respect to an adversary that may be operating both in multiple time periods and platforms. This provides a longitudinal study of attacker behavior, reuse of infrastructure and development of capabilities that are essential in knowing the intent and strategy of the adversary [37]. Entities can also be resolved better and hence less terrain on the false attribution of the false alias reuse or collision of similar names.

The other essential use is in cross platform correlation of campaign. The contemporary cyber campaigns commonly cut across technical infrastructure, social engineering stories and underground communications. The given approach can be seen as a way to rebuild the campaigns holistically by connecting entities in forums, social media, malware repositories, and security reports. This feature improves the detection of early-warmth where the upcoming campaigns can be detected before they are completely developed and facilitates countermeasures [38].

The framework also implicates significant inferences on CTI knowledge graph formation and enrichment. Entity resolution of high quality is a basis of cyber threat knowledge graphs that combine indicators, actors, tools as well as tactics to be built and maintained. Context entwined entity connecting enhances graph consistency and eliminates any redundancy allowing more justified graph analytics e.g., connection inference and attack path analysis [39]. Such enhanced knowledge graphs can directly be incorporated into the Security Operations Centers (SOCs) and automated intelligence platforms.

Organizationally, the enhanced effectiveness of the entity resolution leads to the efficiency of the analysts and decision support. By automating complex tasks of disambiguation load and manual scope can be minimized and the analysts provided with opportunities to reason on higher levels and make strategic evaluations. Additionally, more accurate CTI results aid in to make informed decisions during risk management, incident response prioritization, and information sharing with trusted partners [40].

On the whole, the offered solution facilitates the practical use of artificial intelligence in CTI as it solves a primary problem. Its implementation is capable of improving the precision, scaling, and reliability of OSINT-based intelligence systems that improve cyber defense services in the public and the private sector.

9. Limitations

Although the suggested transformer-based entity resolution framework shows good performance in the Cyber Threat Intelligence (CTI) that is powered by OSINT, it should be noted that several limitations exist. To begin with, the method is based on labeled training data similarity learning, which is both limited and expensive to acquire in CTI fields. Entity relationships grounded on truth are sensitive, unfinished, or may be updated as fresh intelligence keeps coming out. This limitation can restrict the extent to which the model can be generalized to the changing threat environments and new entities that can be seen.

Second, the framework has computational and resource overhead that comes with use of large transformer models. Inference and fine-tuning can be very resource-intensive in terms of processing power and memory, limiting its ability to be deployed in real-time or limit resources, including in edge-based monitoring systems or other smaller security operations centers. This issue can be alleviated by using optimizations on performance and model distillation strategies; however, these methods can decrease accuracy.

Third, the model will depend on the quality and coverage of OSINT sources. OSINT data is always noisy, visible on the surface, biased data, and disproportionately distributed across languages and platforms. This leads to the possibility that entity resolution will differ depending on the region or language and threat ecosystem. Also, the framework mainly addresses textual OSINT and does not comprehensively involve such non-textual cues as network traffic trends, malware binaries, or image-based articles.

The other limitation is with regard to adversarial adaptation. In a bid to avoid detection, threat actors constantly change their strategies and may shift patterns of language usage or may just poison OSINT intentionally feeding it with false data. Although contextual embeddings enhance the robustness, the risk of deception cannot be reduced by only the data-based approach without the human analysis. Lastly, the ethical and legal implications surrounding the large-scale OSINT collection and automated profiling are out of the scope of the present study, yet these factors are essential in respectable deployment.

10. Conclusion

The current paper has described a transformer based entity resolving framework of OSINT-based Cyber Threat Intelligence to deal with ambiguity in alias, discrepant naming and multilingual information. Experimental findings show that it has important advances against the traditional approaches in terms of accuracy and quality of clustering. Through the improvement of cross-source entity linking, the suggested methodology reinforces other fundamental CTI activities including threat correlation and attribution. Although there are some limitations, the research shows that deep learning has a potential of advancing scalable and reliable cyber threat intelligence systems.

References

1. E. M. Hutchins, M. J. Cloppert, and R. M. Amin, “Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains,” Lockheed Martin Corporation, 2011.
2. I. Tounsi and H. Rais, “A survey on technical threat intelligence in the age of sophisticated cyber attacks,” Computers & Security, vol. 72, pp. 212–233, 2018, doi: 10.1016/j.cose.2017.09.001.
3. C. Mavroeidis and S. Bromander, “Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within CTI,” in Proc. 2017 European Intelligence and Security Informatics Conf. (EISIC), 2017, pp. 91–98, doi: 10.1109/EISIC.2017.20.
4. H. Samtani, S. Patton, M. Chinn, C. Chen, and H. Chen, “Entity-centric analysis of cyber threat intelligence reports,” IEEE Computer, vol. 53, no. 3, pp. 66–74, 2020, doi: 10.1109/MC.2019.2950032.
5. V. Husák, J. Kašpar, E. Bou-Harb, J. Čegan, and P. Švenda, “Survey of attack projection, prediction, and forecasting in cyber security,” IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 640–660, 2019, doi: 10.1109/COMST.2018.2871866.
6. J. P. McCarthy, M. A. Maimon, and L. L. Wright, “Malware naming inconsistencies: Measuring the impact of anti-virus label diversity,” in Proc. IEEE Int. Conf. on Intelligence and Security Informatics (ISI), 2016, pp. 154–159, doi: 10.1109/ISI.2016.7745467.
7. T. Rid and B. Buchanan, “Attributing cyber attacks,” Journal of Strategic Studies, vol. 38, no. 1–2, pp. 4–37, 2015, doi: 10.1080/01402390.2014.977382.
8. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
9. J. Saxe and K. Berlin, “Deep neural network based malware detection using two dimensional binary program features,” in Proc. 10th Int. Conf. on Malicious and Unwanted Software (MALWARE), 2015, pp. 11–20, doi: 10.1109/MALWARE.2015.7413680.
10. W. Li, W. Meng, and L. F. Kwok, “Phishing detection using deep learning,” in Proc. 2018 IEEE Int. Conf. on Dependable, Autonomic and Secure Computing, 2018, pp. 1–8, doi: 10.1109/DASC.2018.00008.
11. S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, and V. Arcaute, “Deep learning for entity matching: A design space exploration,” in Proc. 2018 ACM SIGMOD Int. Conf. on Management of Data, 2018, pp. 19–34, doi: 10.1145/3183713.3196926.
12. M. Sabottke, O. Suciu, and T. Dumitraș, “Vulnerability disclosure in the age of social media: Exploiting Twitter for predicting real-world exploits,” in Proc. 24th USENIX Security Symp., 2015, pp. 1041–1056.
13. R. Perdisci, W. Lee, and N. Feamster, “Behavioral clustering of HTTP-based malware and signature generation using malicious network traces,” in Proc. 7th USENIX NSDI, 2010, pp. 391–404.

14. A. Mittal, A. K. Sahu, and H. Jain, “Cyber threat intelligence sharing: An ontology-based approach,” Computers & Security, vol. 87, 2019, Art. no. 101593.
15. Y. N. Silva, B. Kang, and E. Keogh, “Similarity search for web-based cyber threat intelligence,” in Proc. IEEE Int. Conf. on Big Data, 2016, pp. 161–170.
16. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.
17. S. E. Coull and J. Gardner, “Activation analysis of transformer models for cybersecurity text,” in Proc. IEEE Symp. on Security and Privacy Workshops, 2021, pp. 243–250.
18. J. Lin, Y. Wu, and Z. Chen, “Automatic extraction of cyber threat intelligence from unstructured text using deep learning,” IEEE Access, vol. 8, pp. 134092–134103, 2020.
19. H. Samtani, S. Patton, and H. Chen, “Leveraging natural language processing for cyber threat intelligence,” IEEE Computer, vol. 54, no. 3, pp. 72–81, 2021.
20. D. D. S. K. Brown and S. M. Ghaffarian, “Cyber threat intelligence correlation using rule-based and statistical methods,” Journal of Cybersecurity, vol. 6, no. 1, 2020.
21. T. Rid, Active Measures: The Secret History of Disinformation and Political Warfare. New York, NY, USA: Farrar, Straus and Giroux, 2020.
22. Y. Wang, Y. Liu, and J. Zhang, “A survey on deep learning for cybersecurity,” IEEE Access, vol. 8, pp. 184364–184390, 2020.
23. A. Doan, A. Halevy, and Z. Ives, Principles of Data Integration. San Francisco, CA, USA: Morgan Kaufmann, 2012.
24. Y. Liu, J. Guo, and X. Cheng, “Multilingual representation learning for noisy text,” IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 6, pp. 2481–2493, 2021.
25. M. L. Lison and J. Bibal, “Named entity recognition in cybersecurity reports,” in Proc. ACM Workshop on Artificial Intelligence and Security, 2018, pp. 1–10.
26. R. Bommasani et al., “On the opportunities and risks of foundation models,” arXiv preprint arXiv:2108.07258, 2021.
27. K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
28. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD), 1996, pp. 226–231.
29. R. M. Lee, M. Assante, and T. Conway, Analysis of the Cyber Attack on the Ukrainian Power Grid, SANS Industrial Control Systems, 2016.
30. A. McCallum, K. Nigam, and L. H. Ungar, “Efficient clustering of high-dimensional data sets with application to reference matching,” in Proc. 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2000, pp. 169–178.
31. T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in Proc. Int. Conf. on Learning Representations (ICLR), 2013.
32. L. Ertoz, M. Steinbach, and V. Kumar, “Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data,” in Proc. SIAM Int. Conf. on Data Mining, 2003, pp. 47–58.
33. H. Samtani, S. Patton, and H. Chen, “Leveraging natural language processing for cyber threat intelligence,” IEEE Computer, vol. 54, no. 3, pp. 72–81, 2021.
34. T. Rid and B. Buchanan, “Attributing cyber attacks,” Journal of Strategic Studies, vol. 38, no. 1–2, pp. 4–37, 2015.
35. C. Wagner, A. Dulaunoy, G. Wagener, and A. Iklody, “MISP: The design and implementation of a collaborative threat intelligence sharing platform,” Computers & Security, vol. 59, pp. 49–64, 2016.
36. A. L. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.
37. B. G. Carr, J. Ellis, and M. Malan, “Cyber threat actor profiling: A behavioral approach,” IEEE Security & Privacy, vol. 17, no. 5, pp. 72–80, 2019.
38. M. Husák, M. Čermák, T. Jirsík, and P. Celeda, “Survey of attack projection, prediction, and forecasting in cyber security,” IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 640–660, 2019.
39. H. Samtani, S. Patton, and H. Chen, “Entity-centric analysis of cyber threat intelligence reports,” IEEE Computer, vol. 53, no. 3, pp. 66–74, 2020.
40. O. R. Zafer, S. K. Gupta, and A. K. Jain, “Decision support systems for cyber threat intelligence,” Computers & Security, vol. 92, 2020, Art. no. 101741.